Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 27 (2011) 241 – 247

Procedia Social and Behavioral Sciences

Pacific Association for Computational Linguistics (PACLING 2011)

Machine Learning-based Syllabus Classification toward Automatic Organization of Issue-oriented Interdisciplinary Curricula

Susumu Ota[a*] and Hideki Mima[a]

[a]*School of Engineering, The University of Tokyo, 7-3-1 Hongo Bunkyo-ku Tokyo 113-8656, Japan*

Abstract

The purpose of this study is to organize issue-oriented interdisciplinary curricula, in which natural language processing, and machine learning-based automatic classification are combined. The recent explosion in scientific knowledge due to the rapid advancement of academia and society makes it difficult for learners and educators to recognize the overall picture of syllabus. In addition, the growing amount of interdisciplinary research makes it harder for learners to find subjects that suit their needs from the syllabi. In an attempt to present clear directions to suitable subjects, issue-oriented interdisciplinary curricula are expected to be more efficient in learning and education. However, these curricula normally require all the syllabi be manually categorized in advance, which is generally time consuming. Thus, this emphasizes the importance of developing efficient methods for (semi-) automatic syllabus classification in order to accelerate syllabus retrieval. In this paper, we introduce design and implementation of an issue-oriented automatic syllabus classification. Preliminary experiments using more than 850 engineering syllabi of the University of Tokyo show that our proposed syllabus classification system obtains sufficient accuracy.

Education; Machine Learning; Classification; Automatic Term Recognition

* Corresponding author. Tel.: +81-3-5841-8795; fax: +81-3-5841-8917.
E-mail address: ota@t-adm.t.u-tokyo.ac.jp.

1877-7058 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
Selection and/or peer-review under responsibility of PACLING Organizing Committee.
doi:10.1016/j.sbspro.2011.10.604

1. Introduction

With the advances in science and technology in recent years, engineering knowledge has expanded drastically, but on the front lines of engineering education, it has become extremely difficult for students to select courses and exercises suited to their own interests. For example, the University of Tokyo's Faculty of Engineering offers about 900 courses, but not all students have the knowledge required to select courses appropriately. Students should be able to determine which courses a) will teach them the material suited to their interests and b) they should take in preparation for the targeted courses. Also, instructors need to know which materials are redundant or are missing from specified courses in order to increase the effectiveness of the curriculum. It is therefore extremely important to maintain an overview of all courses, as well as to know the relationships among individual courses [4][10].

In this paper, we propose a semi-automatic system for interdisciplinary syllabus classification using an issue-oriented approach to syllabus structuring. The main feature of this system is that in the process of syllabus classification, which in the past was conducted entirely manually, it extracts unique characteristics of those syllabi using language processing, and automatically classifies syllabi on the basis of machine learning, thereby reducing the time required for the process as a whole. We will explain the issue-oriented syllabus classification and structuring system on which our group is working using this issue-oriented syllabus structuring approach. We will also discuss the system configuration, implementation, and experimental evaluation using experimental data.

2. Overview of system

The main purpose of this study is to develop an efficient issue-oriented syllabus retrieval system that presents clear directions to learners. Our approach to developing an issue-oriented syllabus classification system is based on the following:

- Automatic term recognition (ATR) as a feature extraction
- Support vector machine (SVM)-based machine-learning to develop syllabus classifier
- Interaction-based class modification

The system architecture is modular and integrates the following components (Fig. 1):

- *Terminology-based feature extraction (TFE):* carries out the automatic term recognition, which includes term extraction and term variation management.
- *SVM-based leaning:* learns how to classify syllabi by extracting classification patterns from features that are also extracted by TFE and produces classification knowledge.
- *SVM-based classification:* classifies a given syllabus' features by referring to the classification knowledge.
- *Syllabus class visualizer:* visualizes syllabus structures on the basis of HTML expressions in which hyperlinks between class names and syllabi documents are automatically provided (Fig. 2).
- *Interaction-based modifier:* allows users to modify classification errors interactively (Fig. 2).

As the flow in Fig. 1 shows, the system learns classification patterns using manually classified syllabi (e.g. previous years' syllabi and their class information) and produces classification knowledge. Then target syllabi (e.g. this year's syllabi) are classified using the classification knowledge.

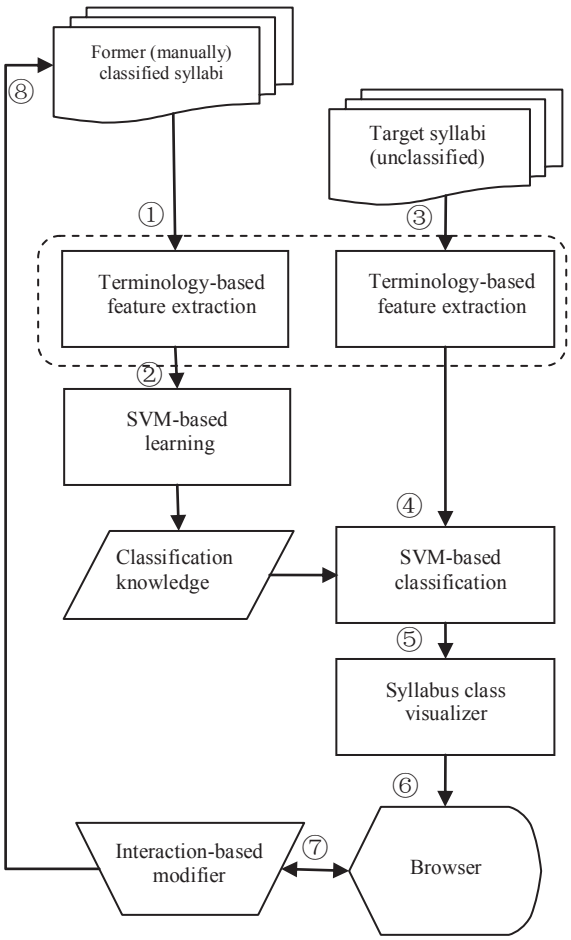

Fig. 1: The system diagram

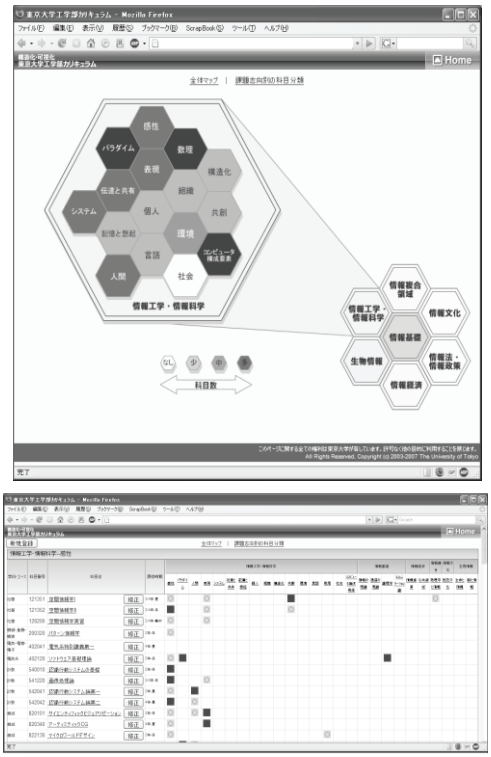

Fig. 2: Sample of classification and visualization of curricula using SVM

3. Support vector machine as machine learning-based automatic classification

The SVM is a powerful pattern matching tool that can be applied to classify input elements in accordance with the past classified patterns. In general, SVM performs classification (categorization) on the basis of the structural risk minimization principle [1][2][9] from computational or machine learning theory. We chose SVM as a classifier to highly accurately classify text. The goal of SVM-based syllabus classification is to classify syllabi into a fixed number of predefined categories by using an SVM learner and SVM classifier. The SVM and its performance have been further detailed elsewhere [1][2][8][9].

The first step in classification is to transform syllabus documents, which normally are texts, into a representation suitable for the learning algorithm and the classification task.

Taira and Haruno reported that in text classification using SVM, the approach to the selection of features is very important [8]. Moschitti and Basili discussed text classification using a variety of features, including n-gram, morphemes, and compound nouns [6]. In that report, the results of a classification experiment using SVM with English newspaper articles showed that when morphemes combined with

compound nouns are used as the features, classification accuracy improved slightly in eight out of ten category types. There are reports that when viewed as a whole, even when features other than "bag of words" (phrases, word meaning information, etc.) are used, the improvements in accuracy are very slight, and that considering the tradeoff with computational complexity, this approach does not necessarily improve quality. Nevertheless, there is still room for discussion on the optimization of features in keeping with the level of specialization in the categories (e.g., credibility in terms of feature quantity in specialized terminology) or the method for combining features. For example, in the field of information searches, in the past, there were discussions about whether to use n-gram or morphemes as the unit for features, but in today's search systems, the accuracy of searches has already been improved by combining both of these elements.

On the basis of the above perspective, when designing this system, we conducted experiments using cross validation of previously learned data and preliminary experiments using Japanese newspaper articles and studied the ways in which adopting nouns and terms as features affected classification accuracy. First, we prepared three types of features for SVM text classification - 1) nouns only; 2) terms only; and 3) nouns and terms - and attached scores using term frequency-inverse document frequency (TF-IDF [7]). We then conducted classification experiments on the newspaper articles in each case. The resulting F values for the experiments overall (harmonic mean of precision and recall rates) were as follows: Nouns only: F=0.6390; 2) terms only: F=0.5462; and 3) nouns and terms: F=0.6560. These results suggest that using nouns and terms as the features achieves the most accurate classification. An even greater improvement in accuracy could be seen in highly specialized categories; for example, in the case of newspaper articles on military matters and international relations. Because syllabi often contain large numbers of specialized terms, we can expect to achieve highly accurate classification by using both nouns and terms as the features. For the reasons outlined above, we have adopted nouns and terms as the features for SVM classification.

4. Terminological processing as a feature extraction in TFE

The lack of clear naming standards in a domain (e.g. biomedicine) makes ATR a non-trivial problem [4]. Also, it typically gives rise to many-to-many relationships between terms and concepts. In practice, two problems stem from this fact: a single term may denote a number of concepts, and, conversely, a single concept may be denoted by more than one term. In other words, some terms have multiple meanings (term ambiguity), and, conversely, other terms refer to the same concept (term variation). Generally, term ambiguity negatively affects information extraction precision, while term variation decreases recall.

These problems highlight the impropriety of using simple keyword-based classification [2]. Obviously, more sophisticated techniques are needed. Such techniques should identify groups of different terms referring to the same (or similar) concept(s), and, therefore, could benefit from relying on efficient and consistent ATR. These methods are also important for organizing domain specific knowledge, as terms should not be treated in isolation from other terms. They should rather be related to one another so that the relationships between the corresponding concepts are at least partly reflected in the terminology.

Terminological processing in our system is carried out on the basis of the C / NC-value method [5] for ATR. Its main purpose is to extract domain-specific terminology as features for SVM-based learning.

Table 1 shows samples of automatically recognized terms using Engineering domain syllabus text of 850 lectures (Faculty of Engineering, University of Tokyo, 2006). As the table shows, the method successfully extracted reasonable and representative terms.

Table 1: Sample of recognized terms

Automatically Recognized Terms	Score
基礎知識 (basic knowledge)	144.55
線形代数 (linear algebra)	77.35
統計力学 (statistical mechanics)	74.00
固体物理 (solid-state physics)	67.20
ベクトル解析 (vector calculus)	65.01
偏微分方程式 (partial differential equation)	62.40
材料力学 (mechanics of materials)	62.13
環境問題 (environmental issues)	60.17

5. Experimentations

In this section, we briefly explain our experimentations for the system, which were conducted to evaluate the quality of the SVM-based syllabus classification component described in section 2 and to analyze feasibility of the system. Note that this paper does not evaluate interactive modification process by users.

5.1. Experimental procedure

About 450 lectures in Faculty of Engineering, University of Tokyo were used in this experiment. Each syllabus was assigned category IDs, each of which represented a category of lectures. In the experiment, we used year 2003's syllabi as a training set (total number of data was 459) and 2006's as testing (total number of data was 446). Over 80% of syllabi changed between 2003 and 2006.

We used SVM software TinySVM [3] as the learning system.

5.2. Experimental results

Fig. 3 shows the precision and recall for each category. Precision is defined as the proportion of correctly classified syllabi in the set of all syllabi returned by classifiers. Recall is defined as the number of correctly classified syllabi as fraction of all must-correctly-classified syllabi. As the figure shows, although the recall is unstable depending on categories, the precision is high enough in almost all categories.

Although it is difficult to justify the method with such unbalanced training sets (for example, some categories only have 1 or 2 training data), we think its practical performance should still be evaluated in the actual environment. Unstableness of recall also seems to be derived from the insufficient number of data in training sets, so we can expect it to improve proportionally as the number of correct training sets increases year by year. Consequently, the system is expected to be practical enough to achieve automatic issue-oriented syllabus classification systems.

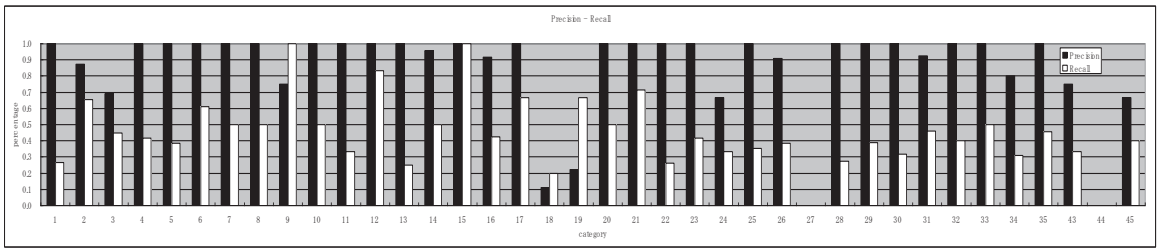

Fig. 3: Precision and Recall of Syllabi Classification Experiments

6. Conclusion

In this paper, we presented a new system for issue-oriented syllabus classification, which offers three main functions: 1) an overview of entire curricula; 2) visualization of the relationships among courses; and 3) visualization of concentration (duplication) and voids in courses. Using this system, it will be possible to automatically categorize large volumes of syllabus data and to interactively correct classification errors, thereby significantly reducing both time and labor costs.

In recent years, with the development of advanced information technologies, society has been evolving at a rapid pace, and has been changing dramatically as a result of factors including the growth of the global economy. The mission of universities is to provide education that adapts quickly to changes in society, but given the demands for constant changes in education - not only through the reorganization of curricula, but through the reorganization of individual subjects as well - the burden of maintaining the educational environment will no doubt continue to grow. In this sense, we believe that there will be steadily increasingly demand for systems like this one, designed to increase the quality of services provided to students, and to facilitate more efficient operations. This system has been experimentally operated on the basis of data from 2003 to 2008 (academic years), and is scheduled to go into full operation soon, providing actual services to students. With the startup of these operations, we will survey students and use their feedback to improve system usability.

References

[1] Cortes, C., and Vapnik, V.. Support Vector Networks, Machine Learning 20:273-297., 1995.

[2] Isbell, C. and Viola, P., Restructuring Sparse High Dimensional Data for Effective Retrieval, Advances in Neural Information Processing Systems, Vol. 11, 1998.

[3] Kudoh, T. "TinySVM." http://www.chasen.org/~taku/software/TinySVM/.

[4] Mima, H, "Structuring and Visualizing the Curricula with MIMA Search." In Proc. 7th APRU Distance Learning and the Internet Conference 2006", 2006.

[5] Mima, H. and Ananiadou, S.. "An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese." International Journal on Terminology, 6 (2), pp. 175-194., 2000.

[6] Moschitti, A. and Basili, R. "Complex Linguistic Features for Text Classification: a comprehensive study." In Proc. 26th European Conference on Information Retrieval Research (ECIR 2004)., 2004.

[7] Salton, G. "Developments in automatic text retrieval." Science, 253, pp. 974-980, 1991.

[8] Taira, H., and Haruno, M., Feature Selection in SVM Text Categorization, Proc. of the Fifth International Conference on Artificial Intelligence (AAAI-99), 480-486, 1999

[9] Vapnik, V., The Nature of Statistical Learning Theory, New York: Springer-Verlag., 1995.

[10] Yoshida, M., "Structuring and Visualizing the Engineering Knowledge -Basic Principles, methods and the application to the UT's Engineering Curriculum." In Proc. 8th APRU Distance Learning and the Internet Conference, 2007.